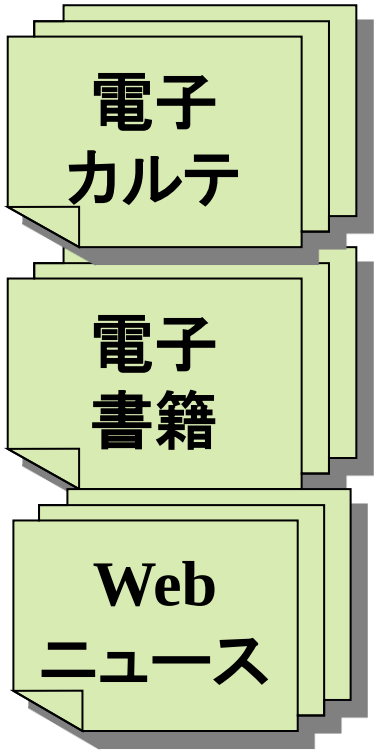
点予測と系列予測の2段階化による品詞推定の精度向上

京都大学 情報学研究科
〇中田 陽介, NEUBIG Graham, 森信介, 河原達也

2011年1月28日（@NHK放送技術研究所）

研究の背景

- 様々な分野で文書の電子化



様々な分野で自然言語処理が要求が高まる

テキストマイニング　評判抽出　自動翻訳　要約・・・

解析処理

- 形態素解析
  - 単語境界推定・品詞推定

⇒様々な分野で高い解析精度が必要

- テキストデータは分野・出典に依存
  - 記述スタイル・出現用語

⇒分野適応性の高い点予測による形態素解析[NL198]

品詞連接の傾向を利用していない！

対象と周辺の文字列を参照する手法

点予測と系列予測の2段階化による
品詞推定の解析精度向上

全体の流れ

点予測による単語境界推定

点予測による品詞推定

提案手法：系列予測による品詞のリランキング

点予測による形態素解析[NL198]

系列予測で利用可能な言語資源

1. フルアノテーションコーパス

   例）|川/名詞|の/助詞|流-れ/名詞|に/助詞|
   |任-せ/動詞|て/助詞|流-れ/動詞|る/語尾|

2. 形態素辞書

   例）川/名詞
   流れ/名詞
   流れ/動詞

点予測でのみ利用可能な言語資源

3. 部分的アノテーションコーパス

例）川の|流-れ|に任せて流れる

川の|流-れ/名詞|に任せて|流-れ/動詞|る

4. 単語辞書

例）鴨川

香川大学

⇒比較的入手・作成が容易な言語資源

⇒高い分野適応性を実現

全体の流れ

点予測による単語境界推定

↓

点予測による品詞推定

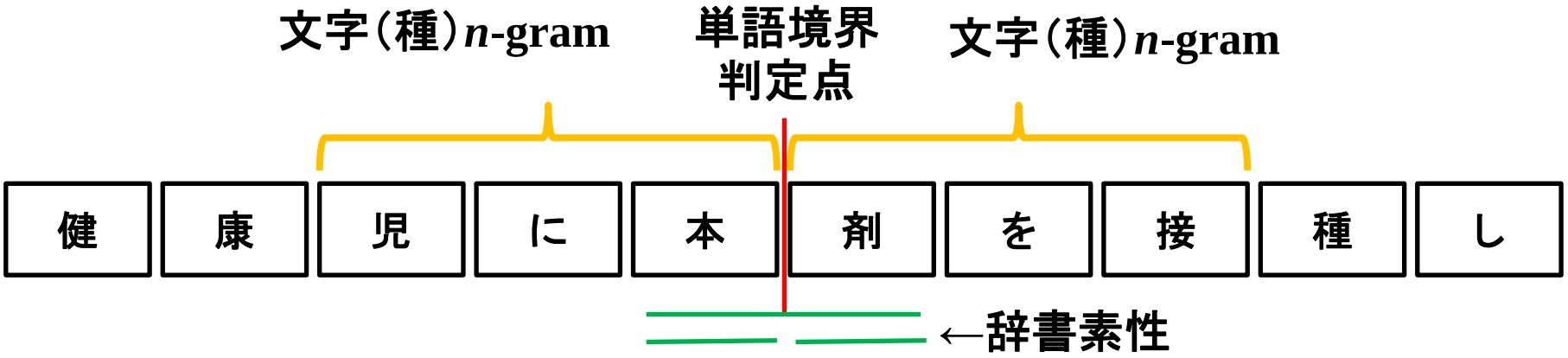
↓

提案手法：系列予測による品詞のリランキング

点予測による形態素解析[NL198]

点予測による単語境界推定

- 単語境界推定は先行研究[Neubigら,NLP10]に従う



  - 点予測:対象と周辺の文字列のみを参照
  - 整形SVMの素性
    - 文字*n*-gram
    - 文字種*n*-gram
    - 単語辞書素性: L1(本), R1(剤), I2(本剤)

全体の流れ

点予測による単語境界推定

点予測による品詞推定

提案手法：系列予測による品詞のリランキング

点予測による形態素解析[NL198]

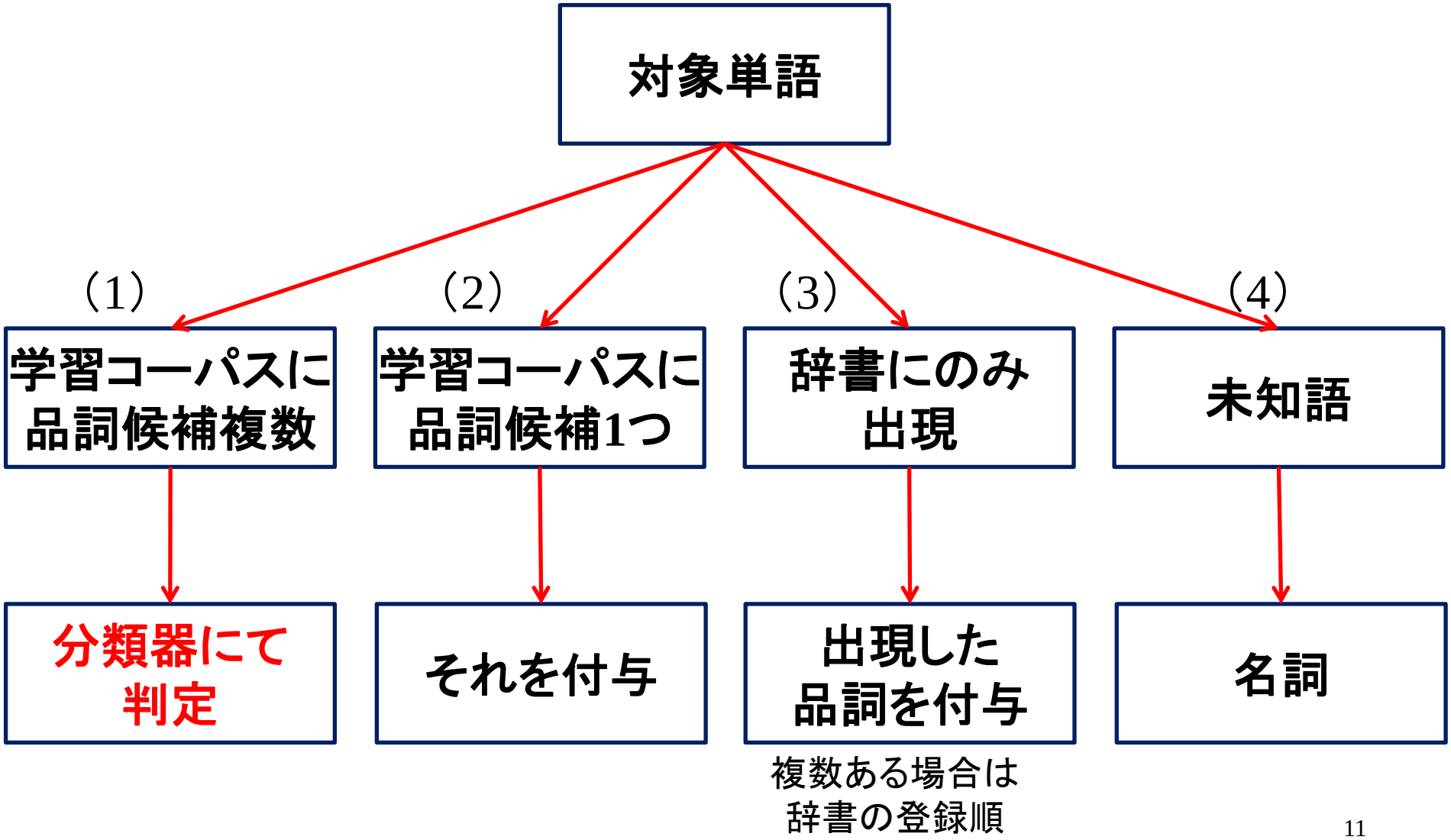
NL198

点予測による品詞推定

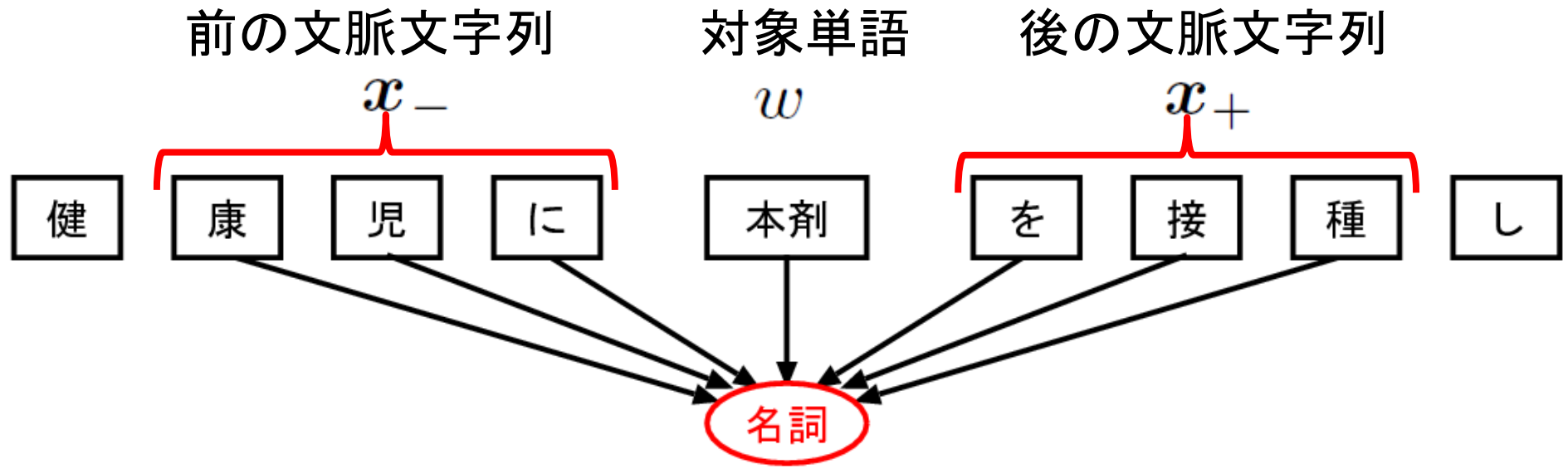
- 点予測による単語境界推定手法を拡張
- 対象単語と周りの文字列を参照

- 柔軟な言語資源の利用
- 高い分野適応性を実現

対象単語により異なる処理



点予測による品詞推定

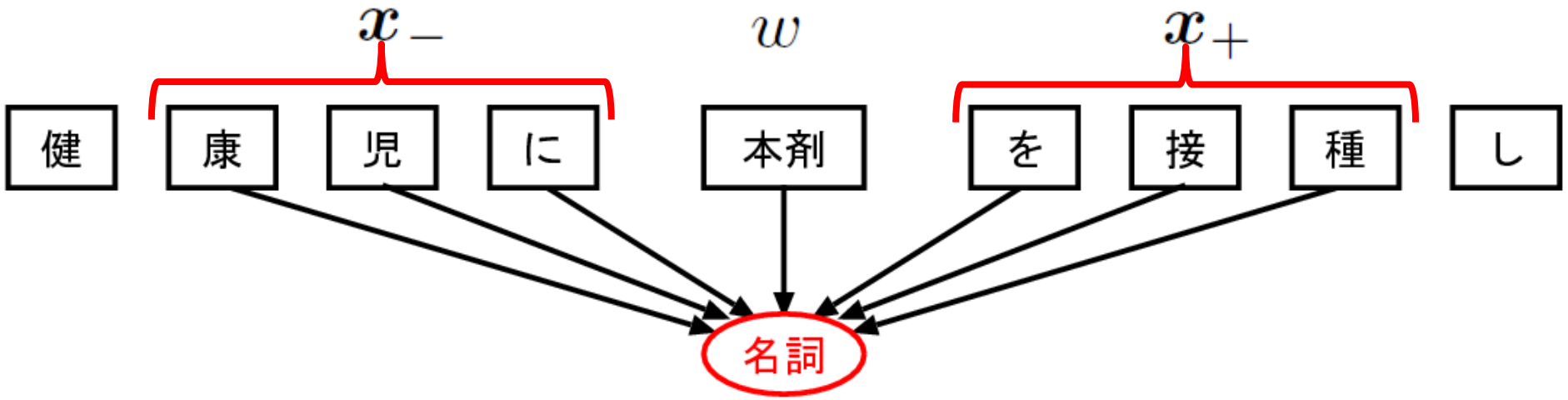


対象単語以外の推定値は参照しない

部分的アノテーションコーパスの利用が可能！

例：食後に　本剤/名詞　の服用

点予測による品詞推定の素性



※ x_- x_+ から成る文字列

- **線形SVMの素性**
  - **文字*n*-gram**
  - **文字種*n*-gram**

柔軟な言語資源の利用

- 部分的アノテーションコーパス

  例）川の|流-れ/名詞|に任せて|流-れ/動詞|る

⇒比較的入手・作成が容易な言語資源

⇒高い分野適応性を実現

全体の流れ

点予測による単語境界推定

↓

点予測による品詞推定

↓

提案手法：系列予測による品詞のリランキング

点予測による形態素解析[NL198]

言語資源の情報

- 単語境界と品詞のフルアノテーションコーパス
  - 単語境界と前後の文字列
  - 形態素と前後の文字列
  - 単語(境界)列
  - **品詞列**
  - 単語/品詞の列

品詞推定において重要な情報源！

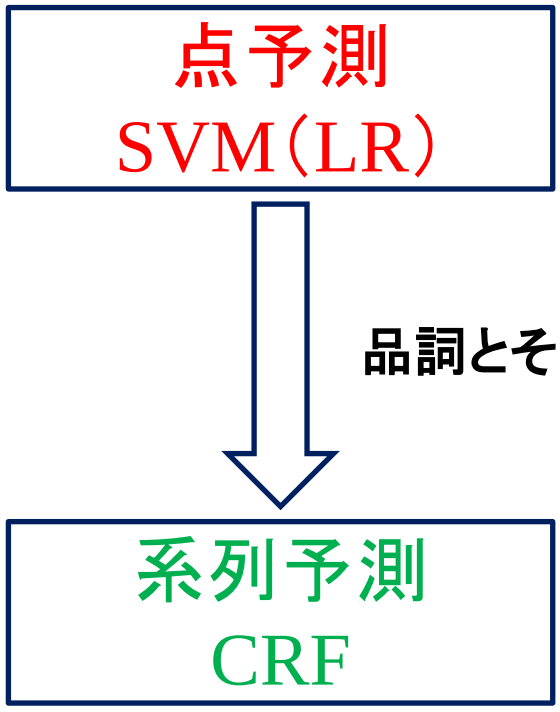
品詞連接の傾向は分野依存性が低く、異なる分野で学習可能と考えられる

⇒品詞連接の傾向を用い、品詞のリランキング

⇒解析精度向上を実現

点予測と系列予測の2段階化

- 点予測と系列予測の2段階化を提案

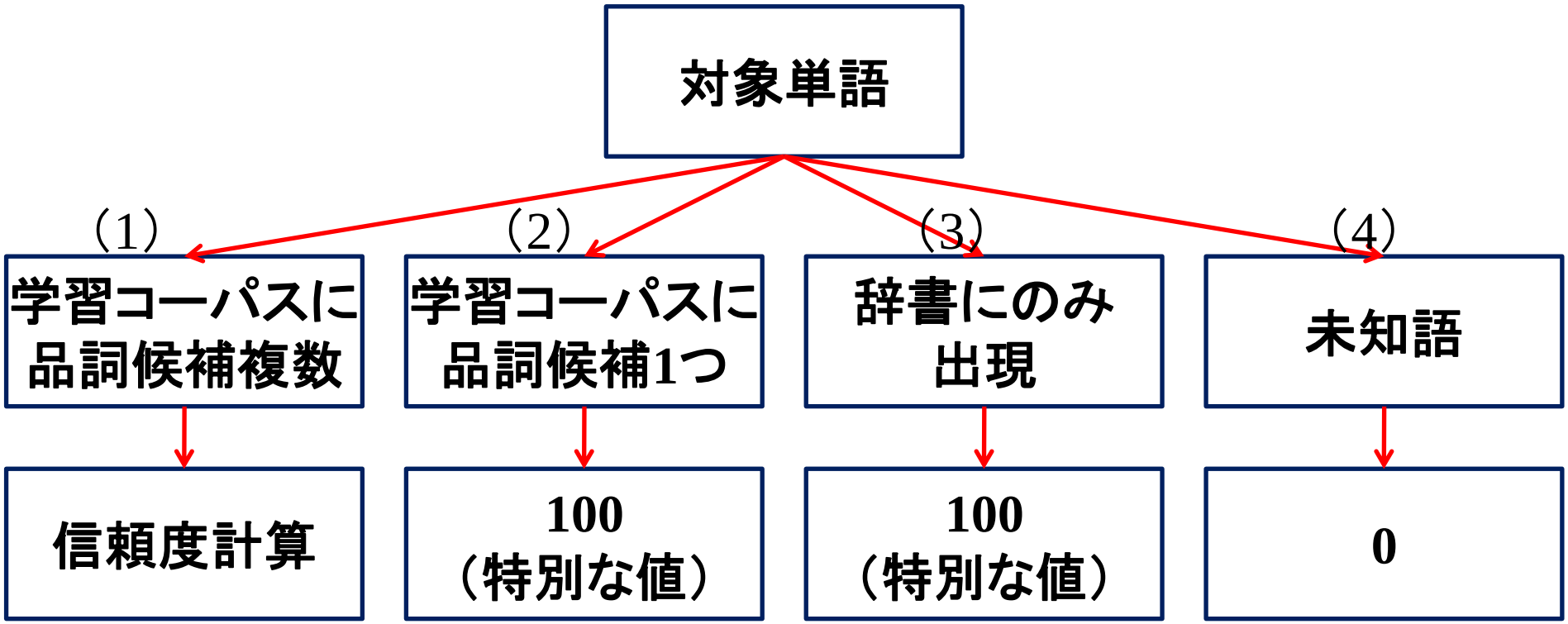


対象以外の推定値を利用しない
部分的アノテーションコーパス利用可能
高い分野適応性追求

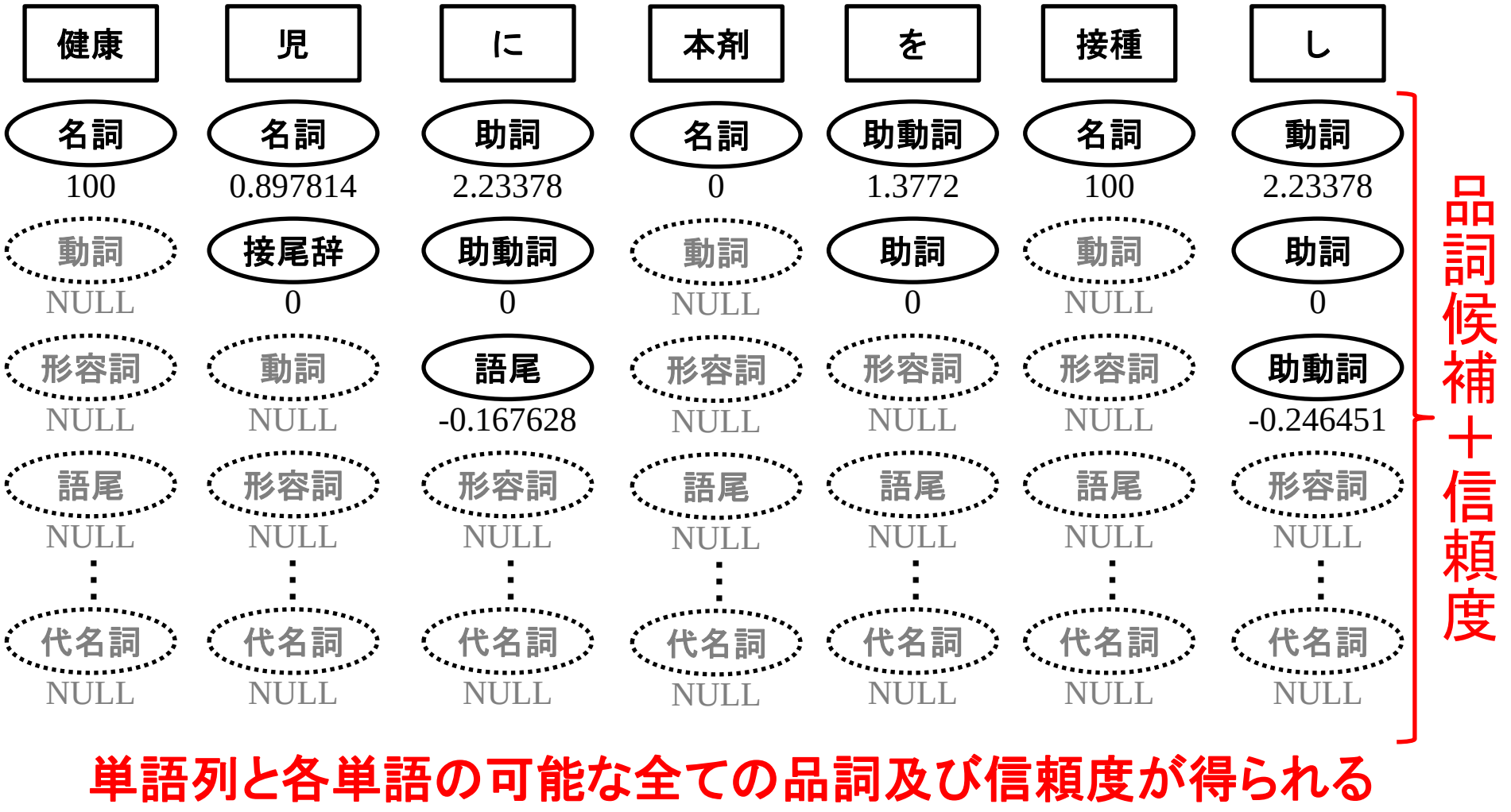
品詞連接の傾向と点予測の結果を利用
文全体で最尤の品詞列を推定
品詞のリランキング
解析精度向上を追求

点予測による信頼度付き品詞推定

- 可能な全ての品詞と信頼度出力
  - 品詞の信頼度は第2候補からの距離を採用



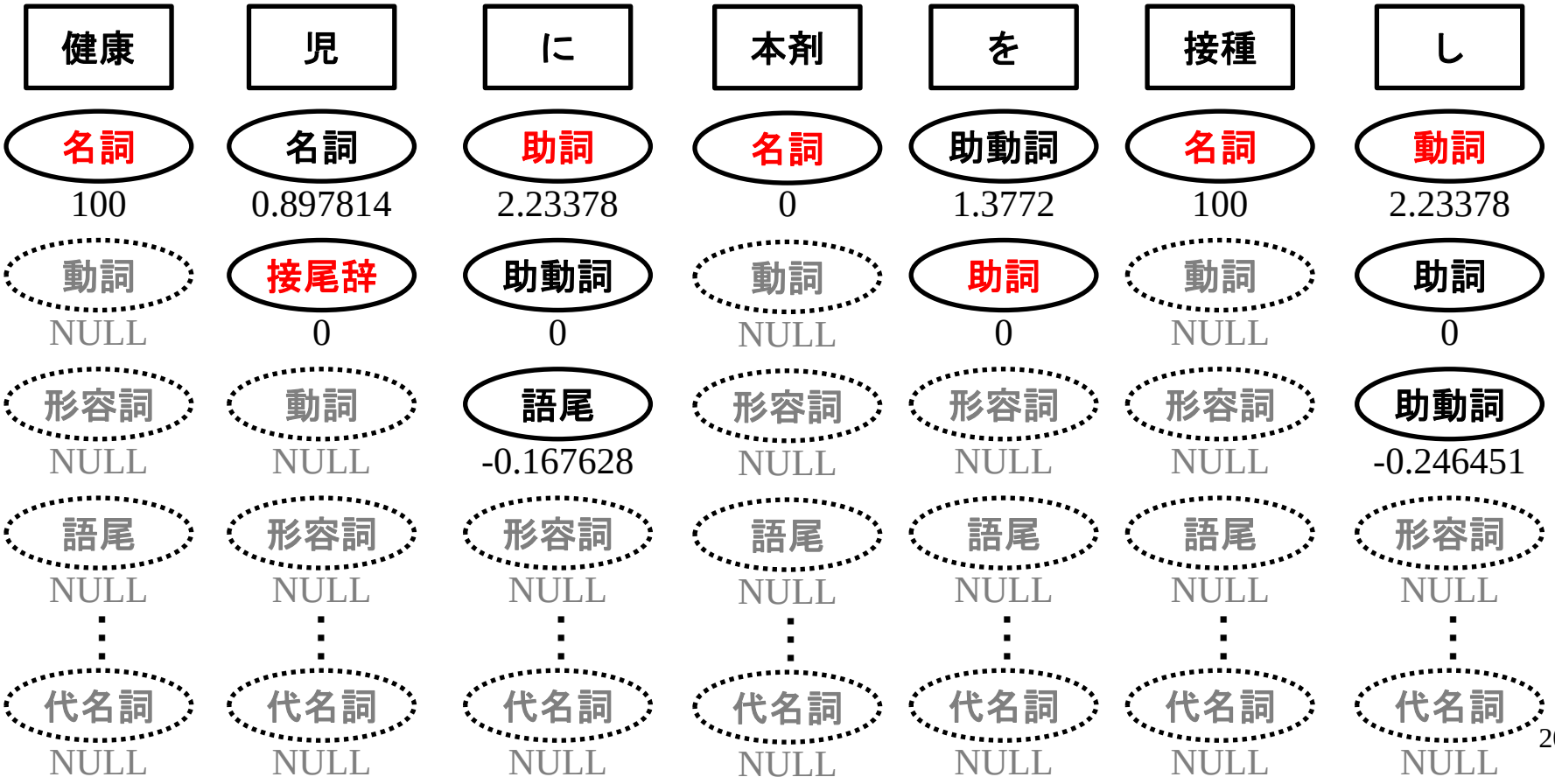
点予測の信頼度付き出力



単語列と各単語の可能な全ての品詞及び信頼度が得られる

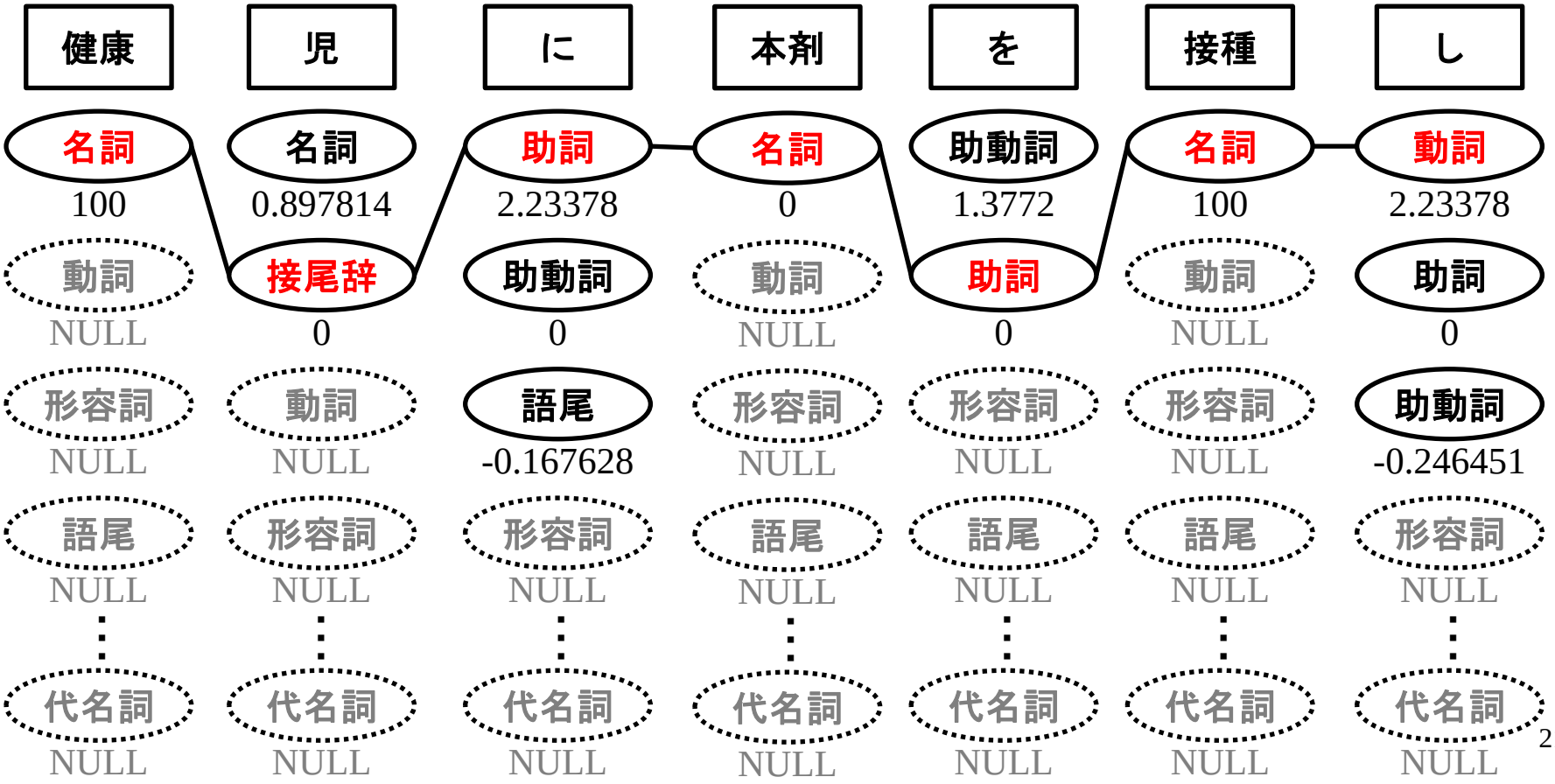
系列予測による品詞のリランキング

- 学習時
  - 入力：正解品詞列
  - 素性：**信頼度素性**と**文脈情報素性**



系列予測による品詞のリランキング

- 解析時
  - 素性：**信頼度素性**と**文脈情報素性**



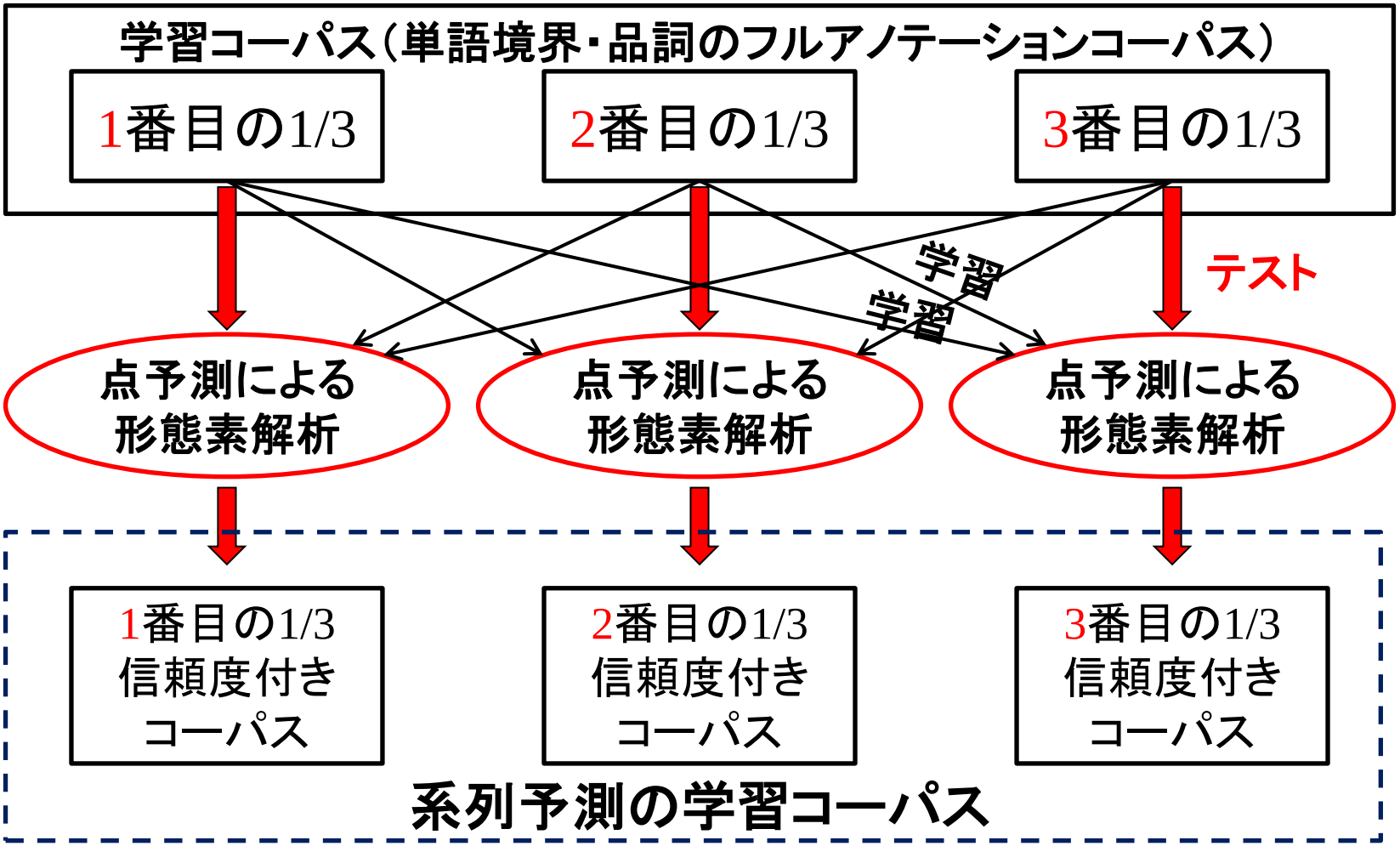
系列予測の素性

※品詞（21種）

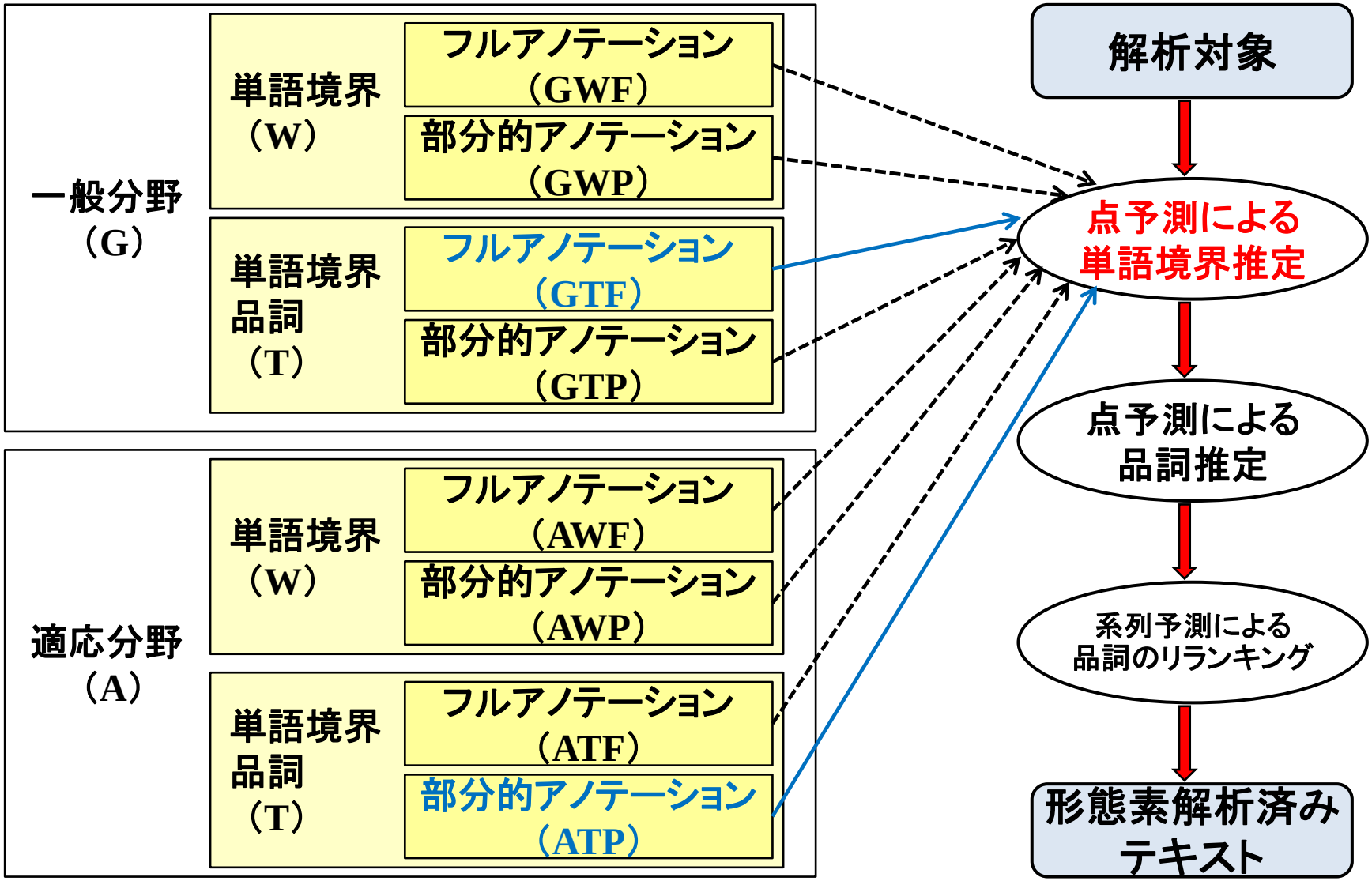
信頼度素性

- 規則1（素性1-21）
  - 信頼度が存在 かつ 100でない場合 その値
  - else NULL
- 規則2（素性22-42）
  - 信頼度が存在しなければ1
  - else NULL
- 規則3（素性43-63）
  - 信頼度が100ならば1
  - else NULL
- 文脈情報素性（素性63-）
  - **単語*n*-gram、文字種*n*-gram**、窓幅5 、*n*の最大値2

系列予測の学習コーパス作成方法

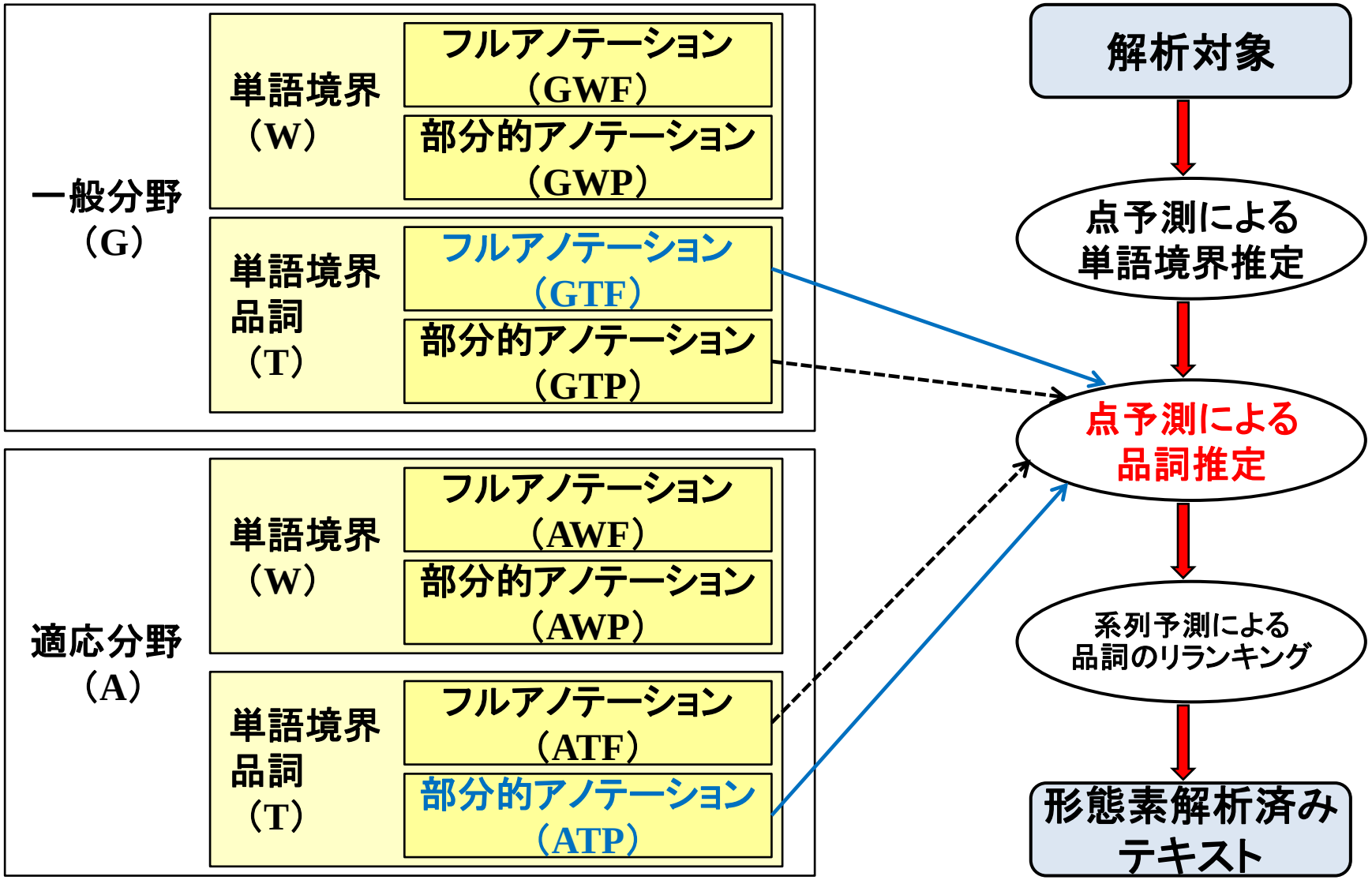


提案する形態素解析の枠組み



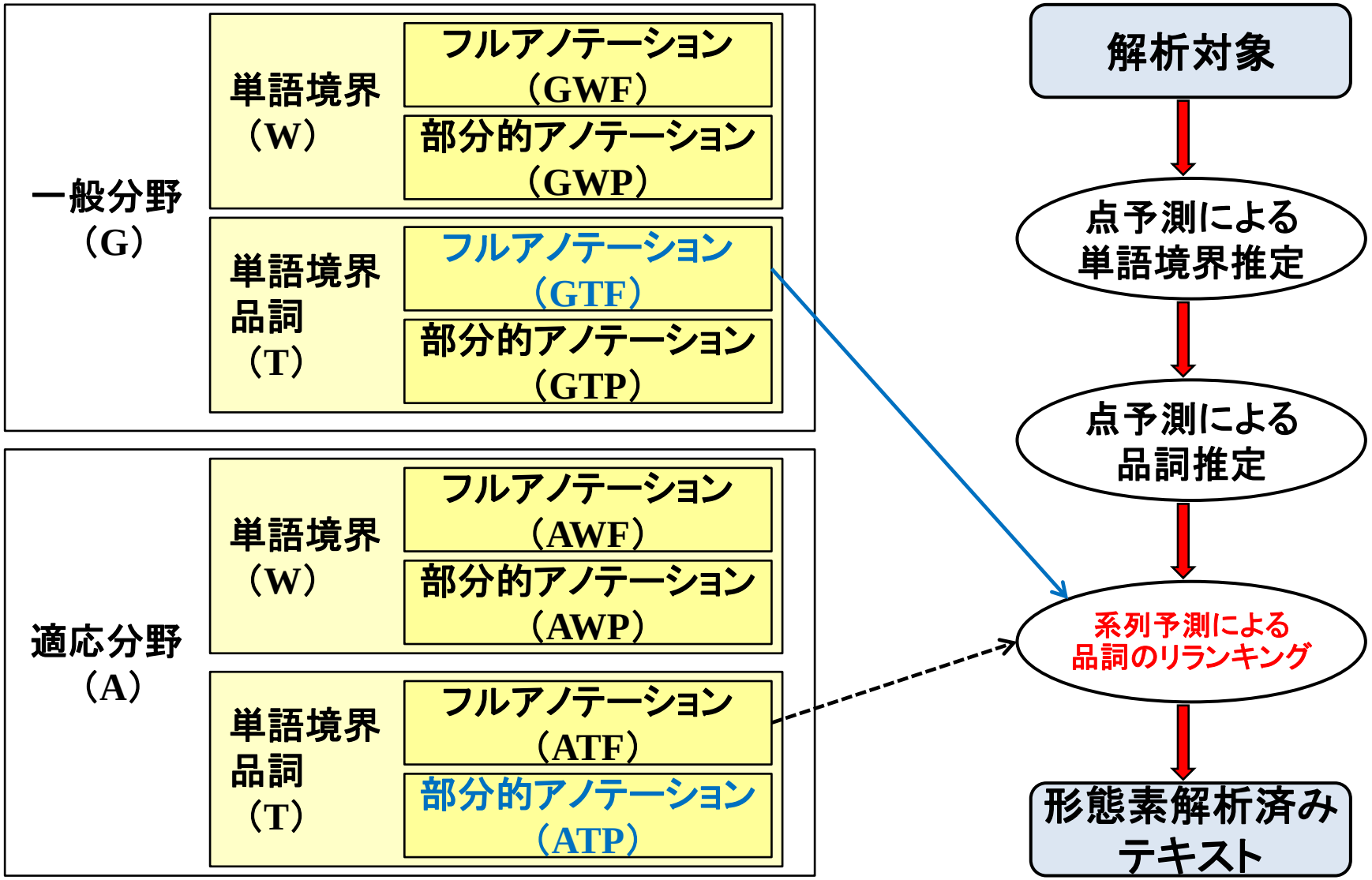
理論的に利用可能なコーパスは破線と実線 の矢印であり、
実験に用いるコーパスは実線の矢印である。

提案する形態素解析の枠組み



理論的に利用可能なコーパスは破線と実線 の矢印であり、
実験に用いるコーパスは実線の矢印である。

提案する形態素解析の枠組み

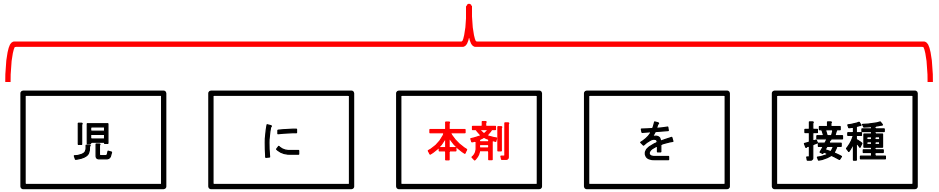


理論的に利用可能なコーパスは破線と実線 の矢印であり、
実験に用いるコーパスは実線の矢印である。

評価実験

- 実験1
  - 提案手法の評価
- 実験2
  - 点予測における分野適応と系列予測による品詞のリランキングの相互作用の評価

- 素性
  - *n*-gramの上限：2
  - 窓幅：5
  - 単語*n*-gram



コーパス

- 日本語書き言葉均衡コーパス（BCCWJ）
  - 単語境界情報
  - 品詞情報（大分類のみ：21種）

出典	用途	文数	形態素数	文字数
白書・書籍・新聞（一般分野）	学習	27,338	782,584	1,131,317
	テスト	3,038	87,458	126,154
Yahoo!知恵袋（適応分野）	学習	5,800	114,265	158,000
	テスト	645	13,018	17,980

評価基準

- 形態素解析精度
  - **正解**：品詞と単語境界が一致した場合正解とする。
  - **再現率**＝（システム正解数／システム出力数）＊100
  - **適合率**＝（システム正解数／正解データ数）＊100
  - **F値**＝（2＊適合率＊再現率）／（適合率＋再現率）

実験1

- 提案手法の評価

- 学習:
  - 一般分野の学習コーパス
    - 系列予測の学習コーパス作成方法に従い作成

- 解析対象:
  - 一般分野のテストコーパス
  - 適応分野のテストコーパス

一般分野に対する形態素解析精度

手法	適合率[%]	再現率[%]	F値（β=1）
品詞2-gramモデル（HMM）	93.77	94.27	94.02
形態素2-gramモデル	96.58	97.65	97.11
形態素3-gramモデル	96.70	97.73	97.21
CRF（MeCab-0.98）	96.72	97.84	97.28
点予測（KyTea-0.1.1）	98.07	98.06	98.06
提案手法	98.41	98.39	98.40

- 提案手法が最も高いF値となった

適応分野に対する形態素解析精度

手法	適合率[%]	再現率[%]	F値（β=1）
品詞2-gramモデル（HMM）	86.78	87.96	87.36
形態素2-gramモデル	92.01	94.09	93.04
形態素3-gramモデル	92.10	94.24	93.16
CRF（MeCab-0.98）	93.69	95.65	94.66
点予測（KyTea-0.1.1）	95.19	95.51	95.53
提案手法	95.86	96.18	96.02

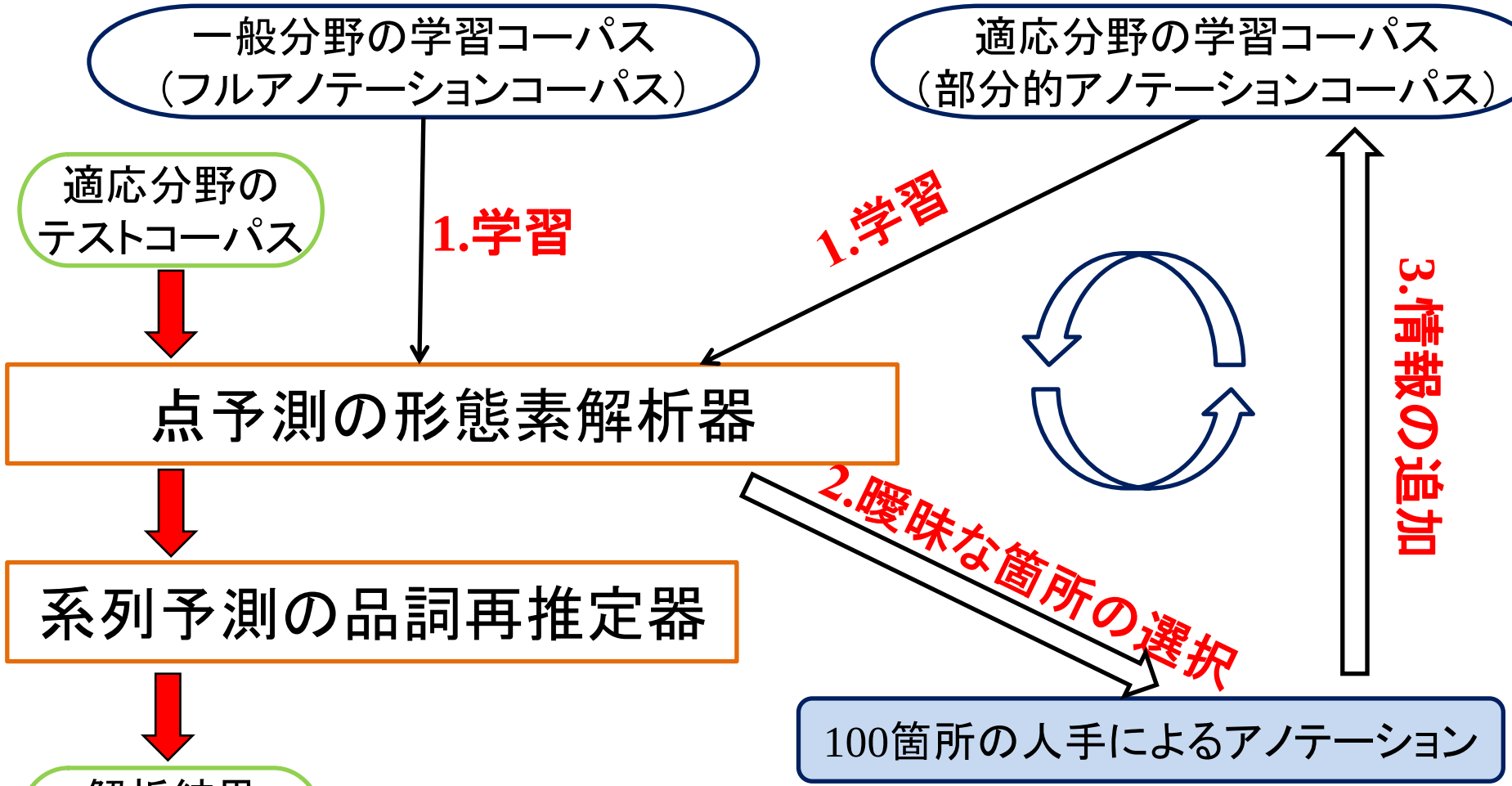
- 異なる分野のテキストに対して、精度は全体的に低下
- 提案手法が最も高いF値を示した
- 品詞連接の傾向は異なる分野からでも学習可能

実験2

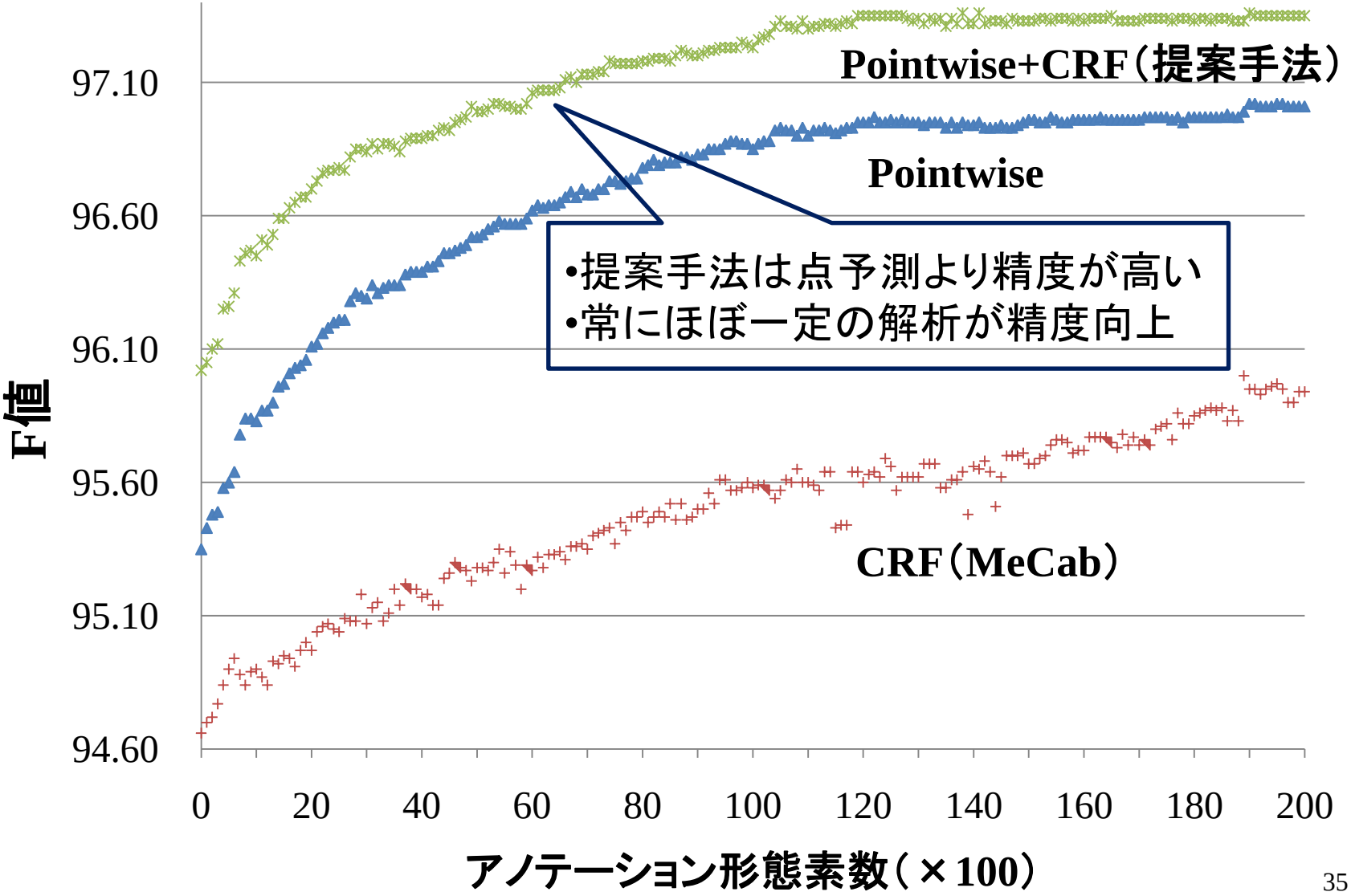
- 分野適応時における品詞のリランキング手法の評価
  - 点予測における分野適応と提案手法の相互作用を評価
  - 部分的アノテーションコーパスを用いた能動学習

- 学習
  - 一般分野の学習コーパス
    - 系列予測の学習コーパス作成方法に従い作成
- 解析対象
  - 適応分野のテストコーパス

部分的アノテーションコーパスを用いた能動学習による分野適応



分野適応結果


Pointwise+CRF（提案手法）
Pointwise
CRF（MeCab）
•提案手法は点予測より精度が高い
•常にほぼ一定の解析が精度向上
F値
97.10
96.60
96.10
95.60
95.10
94.60
0
20
40
60
80
100
120
140
160
180
200
アノテーション形態素数（×100）

まとめ

- 点予測と系列予測による品詞推定手法を提案
  - 点予測による形態素解析の出力を利用
  - 系列予測による品詞のリランキング
  - 解析精度の向上を実現

  - 点予測による分野適応と提案手法の相互作用を評価

  ⇒共に解析精度向上に貢献していることを確認